# Towards Understanding Sycophancy in Language Models

Year: 2023 | Citations: 435 | Authors: Mrinank Sharma, Meg Tong, Tomasz Korbak, D. Duvenaud, Amanda Askell

---

## Abstract

Human feedback is commonly utilized to finetune AI assistants. But human feedback may also encourage model responses that match user beliefs over truthful ones, a behaviour known as sycophancy. We investigate the prevalence of sycophancy in models whose finetuning procedure made use of human feedback, and the potential role of human preference judgments in such behavior. We first demonstrate that five state-of-the-art AI assistants consistently exhibit sycophancy across four varied free-form text-generation tasks. To understand if human preferences drive this broadly observed behavior, we analyze existing human preference data. We find that when a response matches a user's views, it is more likely to be preferred. Moreover, both humans and preference models (PMs) prefer convincingly-written sycophantic responses over correct ones a non-negligible fraction of the time. Optimizing model outputs against PMs also sometimes sacrifices truthfulness in favor of sycophancy. Overall, our results indicate that sycophancy is a general behavior of state-of-the-art AI assistants, likely driven in part by human preference judgments favoring sycophantic responses.